Eur päisches Patentamt
Eur pean Patent Office
Office eur péen des brevets

Bescheinigung Certificate Attestation

Die angehefteten Unterlagen stimmen mit der ursprünglich eingereichten Fassung der auf dem nächsten Blatt bezeichneten europäischen Patentanmeldung überein.

The attached documents are exact copies of the European patent application described on the following page, as originally filed.

Les documents fixés à cette attestation sont conformes à la version initialement déposée de la demande de brevet européen spécifiée à la page suivante.

Patentanmeldung Nr. Patent application No. Demande de brevet n°

02021152.0

Der Präsident des Europäischen Patentamts;
Im Auftrag

For the President of the European Patent Office

Le Président de l'Office européen des brevets
p.o.

R C van Dijk

Anmeldung Nr:
Application no.: 02021152.0
Demande no:

Anmeldetag:
Date of filing: 24.09.02
Date de dépôt:

Anmelder/Applicant(s)/Demandeur(s):

International Business Machines Corporation
New Orchard Road
Armonk, NY 10504
ETATS-UNIS D'AMERIQUE

Bezeichnung der Erfindung/Title of the invention/Titre de l'invention:
(Falls die Bezeichnung der Erfindung nicht angegeben ist, siehe Beschreibung.
If no title is shown please refer to the description.
Si aucun titre n'est indiqué se referer à la description.)

Using a prediction algorithm on the addressee field in electronic mail system

In Anspruch genommene Priorität(en) / Priority(ies) claimed /Priorité(s) revendiquée(s)
Staat/Tag/Aktenzeichen/State/Date/File no./Pays/Date/Numéro de dépôt:

Internationale Patentklassifikation/International Patent Classification/
Classification internationale des brevets:

G06F9/00

Am Anmeldetag benannte Vertragstaaten/Contracting states designated at date of filing/Etats contractants désignées lors du dépôt:

AT BE BG CH CY CZ DE DK EE ES FI FR GB GR IE IT LI LU MC NL PT SE SK TR

DE9-2001-0081

DESCRIPTION

Using a Prediction Algorithm on the Addressee Field in Electronic Mail Systems

1. BACKGROUND OF THE INVENTION

1.1. FIELD OF THE INVENTION

The present invention relates to the field of computer technology, and in particular to a computerized method for predicting the addressee field in an electronic mail system, in which history information associated with the user's sent mail is analyzed for associating the most probable addressee for a given e-mail letter.

1.2. DESCRIPTION AND DISADVANTAGES OF PRIOR ART

When using electronic mail, a user has normally to enter the addressee's complete electronic address. Basically, the user is constrained to enter this address in a correct form, because this address is used in order to route the mail to the intended target mail server. Usually, there is an IP-address of the mail server encoded within the complete addressee's term. Further, user-specific ID-information is comprised of the name, as well, in order to locate a particular addressee out of a plurality of users, registered at the above-mentioned mail server.

Usually, the TCP-IP-address information is not entered by the user, but instead, a better readable character string is used instead. Partly, this character string is converted by the personal e-mail client program into the complete name of a respective person, for instance "David Miller" is shown to the user instead of for instance Miller-David@companyX.de.

This additional information concerning location, organisation and/or country is usually added to the name, in order to make it unique. Sometimes, even a number is concatenated to the name for generating a unique mail system address.

State of the art e-mail client programs installed at the user's PC do not propose or predict an addressee term, i.e. the addressee's e-mail address, when a new mail is created. Once, however, the user begins to enter some characters into the addressee field in the 'new message' window of his e-mail client, the program tries to complete the name based on this initial entry of characters. This holds, for example for Microsoft ™ Outlook Express ™ in one of its latest versions. A similar procedure is known from Microsoft ™ Internet Explorer, Version 6 and many other programs, which append the rest of a stored character string to the initial string entered by a user. The disadvantage is that this automatic string completing needs an initial user action in order to know or to select a significant subset of addressees, which are then selected from a personal address book stored locally at the user PC, and/or stored in a centralised form at the e-mail server in an intranet or LAN of an enterprise.

If the user who creates the new mail does not enter anything into the addressee field, such prior art technique disadvantageously does not make any proposal to whom the mail could be addressed, even if the header / subject line is already filled-in by the user, and/or the text of the mail comprises a considerable number of pages already.

If some initial entry of at least one character is present in the addressee field, a prior art e-mail program, for example the e-mail client in Lotus Notes ™, release 5 performs the above-mentioned proposals based on stored addresses at the mail server of an Intranet, but in order to do that, the user PC must be

disadvantageously connected to the LAN, or to the Intranet, respectively. If the user PC is cut off from such network, only the locally stored address book may be used for addressee term proposals.

Thus, in case of having access to the complete list of electronic mail addressee, i.e., in a "closed world" scenario, where basically all mail addresses are known and can in principle be listed, a prior art e-mail client may do one of the following solutions when confronted with an incomplete or ambiguous mail address entry:

1. it may reject such incomplete or ambiguous mail address;
2. it may replace an incomplete, but unique name in the mail address field on demand or automatically by the complete, unique mail address, or
3. it may map an incomplete and ambiguous mail address to a complete unique mail address and make a respective proposal, in case that more than one complete, unique mail address do match. In prior art usually some heuristic methods are used to rank the list of candidates and then use the top-candidate as the best replacement for the incomplete and ambiguous mail address. Said heuristic methods comprise for example a compare of dates comprised of the history within the e-mail correspondence, or they rely on the number of mails sent to the respective candidates, whereby that candidate is proposed as top candidate who has the most contacts. Thus, some meta data comprised in the history of the user is evaluated in prior art in order to give the best possible addressee term prediction.

Such prediction, however, results today in unsatisfying proposals which make no sense in many cases and may result in misrouted electronic mail.

1.3. OBJECTIVES OF THE INVENTION

It is thus an objective of the present invention to improve the prediction accuracy.

2. SUMMARY AND ADVANTAGES OF THE INVENTION

This objective of the invention is achieved by the features stated in enclosed independent claims. Further advantageous arrangements and embodiments of the invention are set forth in the respective subclaims. Reference should now be made to the appended claims.

In short words, the present invention is based on the idea to apply a combination of prior art Text Mining methods, as for example commercially available in the IBM ™ product "DB2 Intelligent Miner for Text", and prior art Data Mining methods, as for instance commercially available in the IBM ™ product "DB2- Intelligent Mining for Data" , both products being available at each IBM business partner and publicly available under http://www.ibm.com/software/data/iminer/, in order to automatically determine the top favourite addressee in a new e-mail drafted by a user. Thus, basically, prior art Text Mining methods are applied to a set of attributes which are characteristic for e-mail correspondence in order to generate intermediate results in a particular form, preferably in a table-based form, which may be further processed by prior art Data Mining methods. The model generation and training is done once, and then repeated in predetermined intervals, or when it seems appropriate. The application of the generated and trained models is preferably done on-the-fly and transparent to the user, whenever the entry of an addressee is needed. For training and application, a different set of claims is provided, as

training may occur at a place different to that of application, an may be undertaken by different enterprises.

According to its broadest aspect in training mode, a computerized method for predicting the correct addressee to be filled-in an addressee field in a personal electronic mail system is disclosed, in which method user-related history information including the user's sent and/ or received mail is analyzed for associating the most probable addressee for a given e-mail letter. The inventive method is characterized by the steps of:

a) analyzing the contents of at least a subset of the following attributes:
aa) the subject line,
bb) the length of said e-mail letter, or draft, respectively,
cc) the language in use,
dd) the time of day,
ee) the vocabulary in use,
ff) the topics discussed in the body,
gg) the salutation form,
hh) the closing form,
whereby Text Mining methods are used where appropriate for associating attribute values with respective addressees, thus yielding a plurality of single analysis results usable for said prediction,
b) weighting the plurality of said single analysis results in order to provide a Data Mining Model adapted to offer at least one top favorite addressee proposal as a prediction result.

It should be noted that when using the received mail history information, the information from the sender field should be used instead that of the addressee field, which is used in case of doing a training on the sent mail box. Of course, each single information source may be used separately, or they may be used in combination.

Further advantageous embodiments and improvements reveal from the respective dependent claims.

When using separate Data Mining models for different use modes, e.g., office/ private mode of user activities, German/ English language used in the mail, or Confidential/ Not Confidential mail, etc., the underlying models may be selectively trained and prediction quality may be increased.

The same advantage may be obtained when performing a retraining of the user-specific Data Mining model triggered by either of the following criteria:
a) when a user overwrites the addressee proposals made by the e-mail system, more frequently than limited by a predefined threshold level,
b) when the e-mail system is confronted with a number of new addressees, which do not form part of the user's history, and the number or fraction thereof is higher than a predefined threshold level,
c) after a predefined time limit has passed.
Thus, the underlying prediction model is always up-to-date.

When said analysis results are generated in a table-like form, in which each attribute to be analyzed is associated to its predicted value, accompanied by a respective confidence value, then a preferable output form is obtained, which is best worked on subsequently by prior art Data Mining tools, if they are used within the inventive method.

In application mode, according to its broadest aspect a computerized method for completing the addressee field in a user-initiated "new mail" within an electronic mail system is disclosed, which is characterized by the steps of:

a) on occurrence of an incomplete entering of the addressee term

in said addressee field running a predictive Data Mining method based on the trained Data Mining Model as mentioned above,
b) determining at least the most probable addressee to the user as a prediction result.

When further comprising the step of offering a subset of predefined quantity of top favorite addressee proposals to the user, even a respective plurality of similarly assessed top favorites may be proposed to the user for selecting himself the correct single address, or to select further addressees for the above mentioned cc- (copy) or bcc- (blind copy) lists.

When automatically providing an addressee field pre-filled with the top favorite addressee, the case may be well covered by the invention, in which only one very good proposal exists, and the other ones have a significantly worse ranking result.

When testing the Data Mining model on a test set of mails, not being part of the training step, before predicting the most probable addressee, and
issuing a hint to the user, indicating the confidence of the predicted addressee proposal, for instance in form of a confidence value, the user is given additional security for the correct addressee selection, which may save time otherwise needed for a confirmatory check of the personal or Intranet-wide address book.

When further, in case of a trunk of the addressee term being present in the addressee field, e.g., due to manual entering, and in case a high significance of the predictable addressee being present provided by the run of the Data Mining method, said trunk may be automatically expanded with the most probable addressee term. This also contributes to increase user comfort.

When finally, cross-checking a term entered manually by the user with a list of top favorite addressees, determined by the

inventive system, and issuing a warning, if the probability is high that the user-entered term is faulty, then an additional contribution is provided to avoid misrouted mails.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention is illustrated by way of example and is not limited by the shape of the figures of the drawings in which:

Fig. 1, upper part is a schematic block diagram representation showing the basic control flow during generation and training of the prediction models, and in the bottom part during application of said models;

Fig. 2 is a schematic block diagram representation of text mining results used according to the invention.

3. DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

With general reference to the figures and with special reference now to **fig. 1** the basic control flow of the training for the prediction models will be described in more detail.

In a step 110 an inventive program module, which may be incorporated at the personal e-mail client local to the user, or, alternatively in a centralised form at the mail server/firewall, accesses the sent mail and the received mail of a particular user. This "history data" is now subjected according to the invention to prior art Text Mining methodology, step 120. Specifically, a major subset of said history data is used as a training set for predicting the one or more addressees of a mail by exploiting the features, which are specific for e-mail applications, i.e. the following attributes:

The subject line, the salutation and closing sequences, the language or vocabulary in use, and the topics, which are discussed in the body of each e-mail, and optionally any time information available, either in the plain text of the body, or available as meta data in the e-mail client program, and optionally the lengths of the mail, step 130.

The structure of data, which is used in here for analysing the above-mentioned attributes (subject line, etc.) may be selected in a straight forward manner, but preferably in a form in which the succeeding step of processing by a Data Mining method may take profit from. This preferred data structure is illustrated by way of some willingly selected examples in fig. 2.

Further, in a step 140 the plurality of single analysis results resulting from step 130, in which each describes some association between a respective single e-mail and one or more particular addressee candidates thereof, is subjected according to the present invention to a further particular analysis, in order to filter out the less important and less significant attributes which are not very useful when applied for prediction.

According to a basic principle of the present invention and of the preferred embodiment, this further analysis 140 comprises Data Mining based methods, which are in turn and per se known in prior art. The result thereof is depicted by reference sign 150. It is a trained prediction model, which is able to associate a given new e-mail to be created to a quite small set of potential addressees with a probability, which is quite high, at least for the one or two top candidates in a selection list. Said list may be generated by the inventive program module and may be issued to the user for selecting the correct addressee, for instance in case three or four addressees are proposed.

In a further, optional step the trained models 150 may be subjected to a test phase 155, as this is usual in prior art model build and test arrangements. A separate test set of history data, which is preferably disjunctive (has no shared elements) to the training set, is used to predict the addressee, and the system's prediction is compared with the actual addressee. If the rate of correctly predicted addressees is high enough, for instance a success quota of 90 %, the trained models 150 may be used for prediction to be applied for newly created e-mails, thus, some prediction result 170 will be obtained, whenever a user starts writing the plain text of a new e-mail letter, even if he did not fill in any character string into the addressee field.

Once the model is created and optionally tested during the test step 155, the model 150 may be used in its application mode to do one or more of the following:

A first task is to automatically propose an addressee for an e-mail draft 160, which is already written and in which the addressee field is still empty. This may cause the addressee field to be pre-filled. In case that the prediction method in use is able to return a confidence value saying how significant the prediction is, this confidence value may be compared to a threshold and thus help to suppress insignificant predictions. Such threshold levels may be predetermined to comprise several different values dependent upon their use. For example, in private use the thresholds may be lower and in business use they may be higher in order to reflect a gradually increasable care which is necessary to apply in business practice.

Further, the prediction result 170 may be used to automatically expand ambiguous addressee information according to the best match. This is done preferably by giving highest priority to that candidate, who has the highest confidence value. In online mode in case of a closed world scenario, as it was mentioned

above, this addressee term expansion may be visible as a list, from which the user may select a name. In offline mode, result 170 may also be used just to improve the current address resolution in order to avoid sending a mail to a non-intended addressee.

Further, the prediction result 170 may be used for a cross-check between the proposed addressee and the addressee entered manually by the user, in order to raise, i.e. issue a warning, if a predicted addressee is not on the addressee list, and/or if a not-predicted addressee is on the list. In case, the used prediction method is able to return a confidence value, this value can be compared to appropriately selected threshold values, in order to determine, whether either of the above warnings (1. predicted, but not on addressee-list, 2. on addressee-list, but not predicted) should be raised.

With reference now to **fig. 2** a sample set of attributes is described along with the data structure used in the inventive context, required for Data Mining methods. The data structures given in fig. 2 may be seen as a sample, which is hold expressedly simple. They result from the Text Mining method, see back to step 120 in fig. 1. The set 130 of attributes provided by said intermediate Text Mining result (fig. 1) may have a table-like form as depicted in fig. 2 and may comprise a plurality of attributes each accompanied by a confidence value, for example expressed as a percentage.

Preferably, the following attributes are evaluated according to the present invention:

The subject line showing some header topic 210, which gives the reference (RE) may be analysed by Text Mining, in order to retrieve similarities between different topics present in the training data.

Further, the salutation and/or closing attributes 215 may be analysed, up to which degree those attributes are more or less formal. The salutation 'Hello Joe' would be associated for instance with a confidence value of only 5 % to be formal, whereas a salutation 'Dear Sirs' would be assessed with a confidence value of nearly 100 %. Similar rules apply for intermediate degrees, like in cases as 'Hello Mr. Miller' etc..

In the same way the language style 220, i.e., the vocabulary used in the body of the e-mail may be analysed to be more or less formal. Hereby, slang-type wordings contribute to the language to be classified as not formal, whereas the absence of such slang generates a confidence value of high degree, corresponding to a formal style of language.

Further, the language may be analysed by Text Mining methods, in order to reveal a confidence value for a predetermined selection of foreign languages. In fig. 2 only one language 230 is depicted. It should, however be understood that for each language of this selection a respective confidence value may be generated. When the text only comprises German words, for example, the confidence value of German would be 100 %. In a mixed-language case the confidence value may vary according to the proportion of the respective different languages words. In this case, a subset 230 of for example ten different languages may be subjected to Text Mining methods.

According to the invention, a further attribute 240 concerns the confidentiality of an e-mail. According to the invention, an existing e-mail may be trained to be confidential, if one or more of predetermined keywords exist either in the subject line, or in the body of the text. Further, other meta information may be used, in order to assess a mail as confidential. An example is, if the mail has an attachment, which is present in an encoded form.

Further, the text body 250 may be associated with one or more predetermined categories, some of which are given in the bottom box of fig. 2 as an example only, such as politics, sports, economy, project 1, sale of product x, project 2, management of congress x, project 3, sale of product y, project 4, test of product z.

Each of those topics is again accompanied by a percentage field which is provided to store the confidence value calculated by the applied text mining method. Thus, a given e-mail may achieve a score of 30 % politics, 20 % sports and 50 % sale of product x, which corresponds to project 1. The Text Mining analysis done in context of this analysis corresponds completely to prior art.

The inventional method may further be implemented in a way, which provides one prediction model per user. This approach may be, however, varied because there may be cases, where a group of users may share the same prediction model, as they share the same contacts. This helps to simplify and standardise the prediction. Further, the inventional method may be implemented to cover the usually prevailing three different types of addressee lists, as are "to", "cc:", and "bcc". It may be used for either list separately, and/or may be used for the merged contends of some or all lists.

Instead for determining the most probable addressee of a freshly drafted e-mail, the inventional concepts can also be applied for determining, if such e-mail letter should be sent as "CONFIDENTIAL INFORMATION" or not, thus implying respective encoding techniques, or not. For this purpose, a respective training phase can be run through, by the aid of which a respective model is established, which helps to decide, if a document is confidential or not. For instance, such training may generate a predetermined list of keywords, and the letter draft is scanned for such words, for example during the Text Mining processing of "topics", as described above. If one or more of

such keywords occur, the user may be alarmed that the mail should be sent according to the rules usually applied for confidential information.

The present invention can be realized in hardware, software, or a combination of hardware and software. A prediction tool according to the present invention can be realized in a centralized fashion in one computer system or in a distributed fashion where different elements are spread across several interconnected computer systems. Any kind of computer system or other apparatus adapted for carrying out the methods described herein is suited. A typical combination of hardware and software could be a general purpose computer system with a computer program that, when being loaded and executed, controls the computer system such that it carries out the methods described herein.

The present invention can also be embedded in a computer program product, which comprises all the features enabling the implementation of the methods described herein, and which - when loaded in a computer system - is able to carry out these methods.

Computer program means or computer program in the present context mean any expression, in any language, code or notation, of a set of instructions intended to cause a system having an information processing capability to perform a particular function either directly or after either or both of the following

a) conversion to another language, code or notation;

b) reproduction in a different material form.

C L A I M S

1. A computerized method for predicting the correct addressee to be filled-in an addressee field in a personal electronic mail system, in which method user-related history information including the user's sent and/ or received mail is analyzed for associating the most probable addressee for a given e-mail letter,
characterized by the steps of:

a) analyzing (120) the contents of at least a subset of the following attributes:
aa) the subject line (210),
bb) the length of said e-mail letter, or draft, respectively,
cc) the language (230) in use,
dd) the time of day,
ee) the vocabulary (220) in use,
ff) the topics (250) discussed in the body,
gg) the salutation (215) form,
hh) the closing (215) form,
whereby Text Mining methods are used (120) where appropriate for associating attribute values with respective addressees, thus yielding a plurality of single analysis results (130) usable for said prediction,
b) weighting (140) the plurality of said single analysis results in order to provide a Data Mining Model (150) adapted to offer at least one top favorite addressee proposal as a prediction result.

2. The method according to claim 1, further comprising the step of using separate Data Mining models for different use modes.

3. The method according to claim 1, further comprising the step of:

performing a retraining of the user-specific Data Mining model triggered by either of the following criteria:
a) when a user overwrites the addressee proposals made by the e-mail system, more frequently than limited by a predefined threshold level,
b) when the e-mail system is confronted with a number of new addressees, which do not form part of the user's history, and the number or fraction thereof is higher than a predefined threshold level,
c) after a predefined time limit has passed.

4. The method according to claim 1, in which the analysis results are generated in a table-like form, in which each attribute to be analyzed is associated to its predicted value, accompanied by a respective confidence value.

5. A computerized method for completing the addressee field in a user-initiated "new mail" (160) within an electronic mail system, characterized by the steps of:

a) on occurrence of an incomplete entering of the addressee term in said addressee field running a predictive Data Mining method based on a trained Data Mining Model (150) according to claim 1,
b) determining at least the most probable addressee to the user as a prediction result (170).

6. The method according to claim 5, further comprising the step of:
offering a subset of predefined quantity of top favorite addressee proposals to the user.

7. The method according to claim 5, further comprising the step of automatically providing an addressee field pre-filled with the top favorite addressee.

8. The method according to claim 5, further comprising the step of testing (155) the Data Mining model on a test set of mails, not being part of the training step before predicting the most probable addressee, and issuing a hint to the user, indicating the confidence of the predicted addressee proposal.

9. The method according to claim 5, further comprising the step of:
in case of a trunk of the addressee term being present in the addressee field, and in case a high significance of the predictable addressee being present provided by the run of the Data Mining method,
automatically expanding said trunk with the most probable addressee term.

10. The method according to claim 5, further comprising the step of:
cross-checking a term entered by the user with a list of top favorite addressees, determined by the system, and issuing a warning, if the probability is high that the user-entered term is faulty.

11. A computer system having means for performing the steps of a method according to one of the preceding claims 1 to 4, or 5 to 10.

12. A computer program for execution in a data processing system comprising computer program code portions for performing respective steps of the method according to anyone of the preceding claims 1 to 4, or 5 to 10, when said computer program code portions are executed on a computer.

13. A computer program product stored on a computer usable medium comprising computer readable program means for

causing a computer to perform the method of anyone of the claims 1 to 4, or 5 to 10, when said computer program product is executed on a computer.

ABSTRACT

Using a Prediction Algorithm on the Addressee Field in Electronic Mail Systems

The present invention relates to the field of computer technology, and in particular to a computerized method for predicting the addressee field in an electronic mail system, in which method user-related history information including the user's sent and/ or received mail, is analyzed for associating the most probable addressee for a given e-mail letter. In order to improve the prediction accuracy it is proposed to:
a) analyze (120) the contents of at least a subset of the following attributes:
aa) the subject line (210),
bb) the length,
cc) the language (230) in use,
dd) the time of day,
ee) the vocabulary (220), in use,
ff) the topics (250) discussed in the body,
gg) the salutation form (215),
hh) the closing (215),
whereby Text Mining methods are used (120) where appropriate for associating attribute values with respective addressees, thus yielding a plurality of single analysis results (130) usable for said prediction,
b) weight ((140) the plurality of said single analysis results in order to provide a Data Mining Model (150) adapted to offer at least a top favorite addressee proposal (170). (Fig. 1)

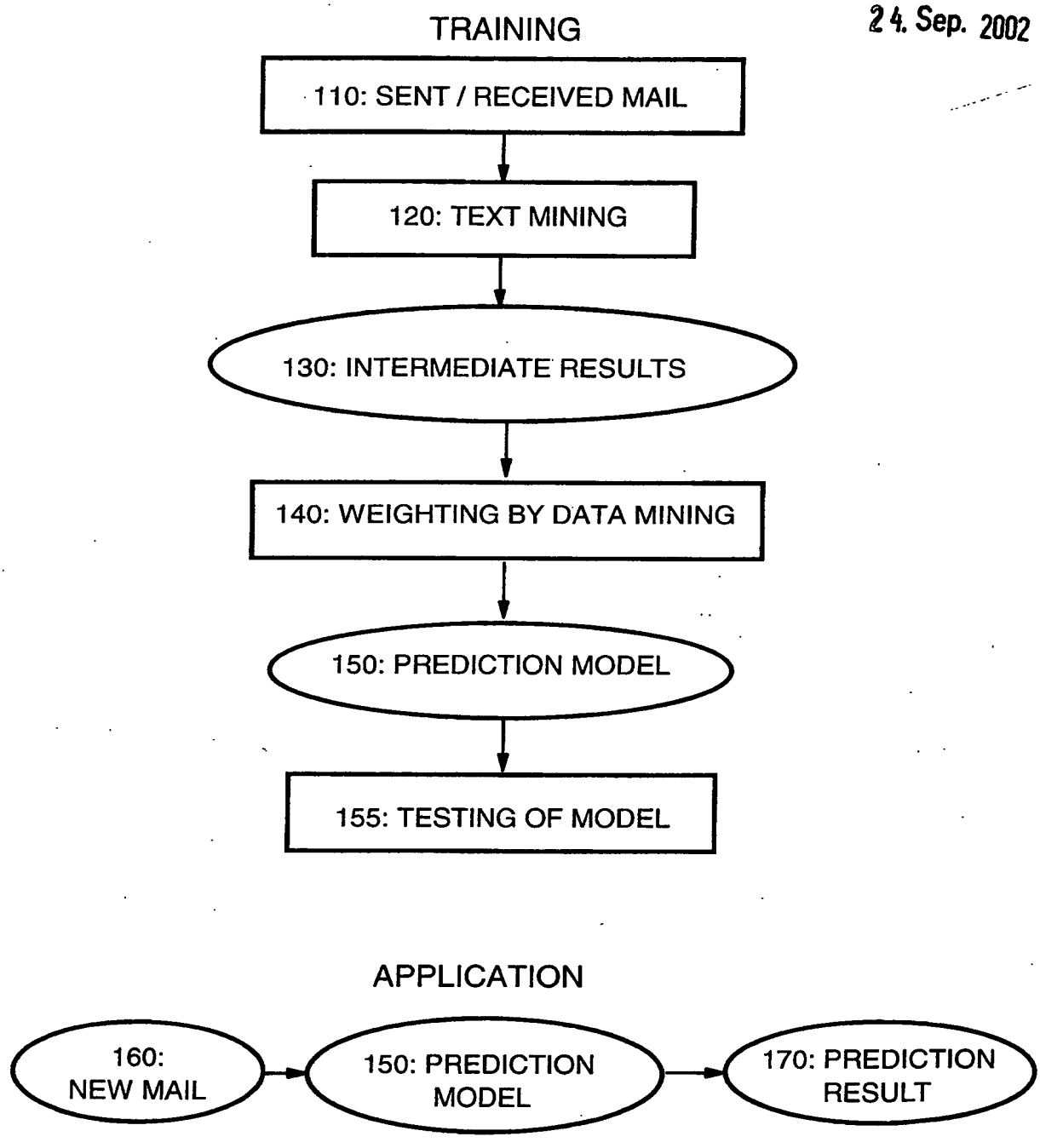

FIG. 1

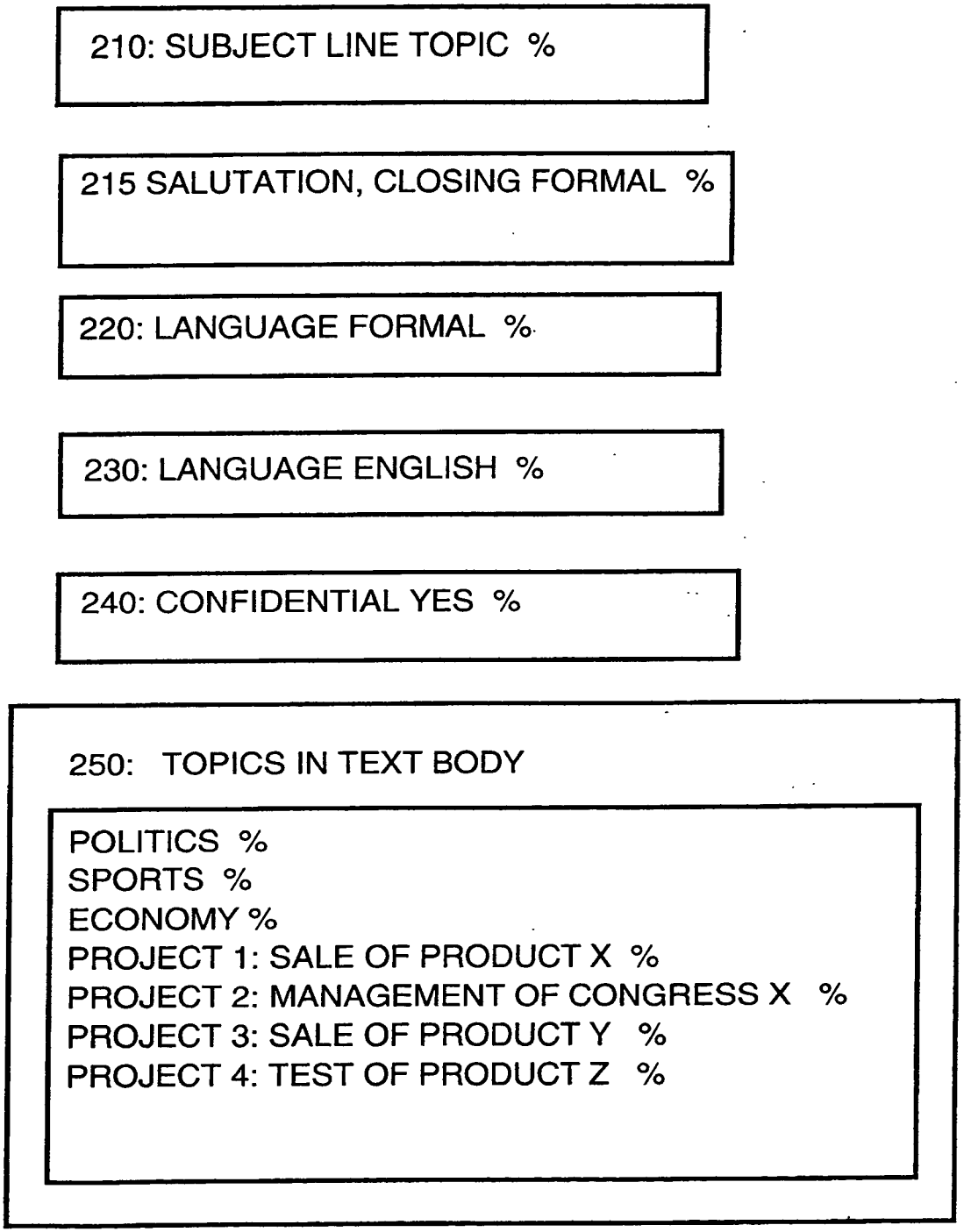
210: SUBJECT LINE TOPIC %
215 SALUTATION, CLOSING FORMAL %
220: LANGUAGE FORMAL %
230: LANGUAGE ENGLISH %
240: CONFIDENTIAL YES %
250: TOPICS IN TEXT BODY
POLITICS %
SPORTS %
ECONOMY %
PROJECT 1: SALE OF PRODUCT X %
PROJECT 2: MANAGEMENT OF CONGRESS X %
PROJECT 3: SALE OF PRODUCT Y %
PROJECT 4: TEST OF PRODUCT Z %

FIG. 2